DOCUMENT RESUME

ED 092 585 TM 003 696

AUTHOR Reckase, Mark D.
TITLE An Application of the Rasch Simple Logistic Model to Tailored Testing.
PUB DATE 74
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Ability; Complexity Level; *Computer Assisted Instruction; Intelligence Tests; Measurement Techniques; *Models; *Multiple Choice Tests; Response Style (Tests); Test Construction; *Testing; Test Reliability
IDENTIFIERS Rasch Simple Logistic Model; *Tailored Testing

ABSTRACT

An application of the two-paramenter logistic (Rasch) model to tailored testing is presented. The model is discussed along with the maximum likelihood estimation of the ability parameters given the response pattern and easiness parameter estimates for the items. The technique has been programmed for use with an interactive computer terminal. Use of the procedure is described in a flexible achievement testing setting. Results are presented showing the number of items needed for good estimation. The independence of items used and ability estimation is shown. Applications of the system to intelligence testing are discussed. (Author)

AERA 27.18

ED 092585

TM 003 696

An Application of the Rasch Simple Logistic Model to Tailored Testing*

by

Mark D. Reckase

University of Missouri-Columbia

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

I. Introduction

Tailored testing (Lord, 1970) can be defined as an evaluation procedure that attempts to administer a test to an individual composed only of items of an appropriate level of difficulty and only as many items as are needed for the purpose of the test. Most currently used paper and pencil tests do not meet these specifications. A fixed set of items is administered to every individual regardless of whether the items are too hard or too easy and every person takes every item. Ideally, a tailored testing procedure would select items for each individual from a large item pool, possibly administering a different set of items to each individual. A preset stopping rule would terminate a testing session when enough information had been gained on an individual, possibly administering a different number of items to each individual.

Tailored testing procedures have made their appearance mainly in response to problems with conventional testing situations. These problems include: inefficient use of examinee time, limited test feedback, improper level of item difficulties, administrator variables, answer sheet effects, time limits and many others.

* Paper presented at the 1974 annual meeting of the American Educational Research Association.

BEST COPY AVAILABLE

Reviews of the literature by Linn, Rock, and Cleary (1969) and Weiss and Betz (1973) give extensive coverage to these problems so they will not be discussed here. Since tailored testing procedures generally are untimed, can give immediate feedback, give items of appropriate difficulty level and solve many other testing problems, they seem the obvious solution to difficulties in the traditional test setting. Current research in the area has yielded many different kinds of tailoring procedures under as many different names. The purpose of this paper is to present yet another tailoring procedure along with a computer program for its implementation. The capabilities and limitations of the procedure will be explored.

The model chosen as a basis for the tailored testing procedure presented here is the Rasch simple logistic model (Rasch, 1960). This model is thought to be a natural choice for tailored testing because of its simplicity, the estimation of the ability parameters independent of the item parameters, and the estimation of item parameters independent of the sample. These properties allow items to be calibrated on various different groups yielding comparable results, and then using a different set of calibrated items to estimate each individual's ability while still yielding ability estimates on the same scale. The details of the procedure will be presented in Section II of this paper.

The actual implementation of the tailored testing employs the capabilities of a time-sharing computer system. Through the use of computer terminals test items are administered, and an interactive computer programs has been written to select items and estimate ability parameters as the administration is taking place.

BEST COPY AVAILABLE

BEST COPY AVAILABLE

II. Theoretical Framework

The tailored testing procedure presented in this paper is based on the Rasch simple logistic model (Rasch, 1960). This model states that the probability that person s will get item i correct is a function of two parameters: the ability of the person, A_s, and the easiness of the item, E_i. More specifically

$$P\{X_{si}\} = \frac{(A_s E_i)^{X_{si}}}{1 + A_s E_i}, \qquad X_{si} = 0,1,$$

where $X_{si} = 1$ if the item is answered correctly and $X_{si} = 0$ if the item is answered incorrectly. Both parameters, A_s and E_i, range from 0 to ∞. If $A_s = 0$, person s has no ability and obviously the probability of a correct response will be zero. As A_s approaches infinity, the probability of a correct response approaches 1.0. If $E_i = 0$, the probability of a correct response is zero and therefore item i is extremely difficult. As E_i approaches infinity, the probability of a correct response approaches 1.0 and hence the item is extremely easy.

This model is a special case of the general three item parameter logistic model developed by Birnbaum (1968). The three parameter model is given by the following formula

$$P\{X_{si} = 1\} = c_i + \frac{(1 - c_i)e^{a_i(\theta_s - b_i)}}{1 + e^{a_i(\theta_s - b_i)}},$$

BEST COPY AVAILABLE

where c_i is a guessing parameter, a_i is a discrimination parameter, b_i is a difficulty parameter, and θ_s is the ability parameter for individual s. The simple logistic model can be obtained from the above formula by setting $c_i = 0, e^{\theta_s} = A_s$, $e^{b_i} = E_i$, and $a_i = 1.0$.

From the relationship of the simple logistic model to the three parameter logistic model, two of the assumptions of simple model can easily be determined. First, the simple logistic model assumes that the probability of a correct response by guessing is zero ($c_i = 0$), and second, all of the items are assumed to be of equal discrimination ($a_i = 1$). Neither of these assumptions is actually justified in practice. Multiple choice items are used for the tailored testing procedure so there is a guessing probability and unless items are selected very carefully there will be some variation in item discrimination. However, Ross (1966) has found that guessing has little effect on the Rasch model and Panchapakesan (1969) and Hambleton (1969) have shown that some variation in the discrimination parameter will not affect the fit of the model.

Two other assumptions also need to be made when using the Rasch model. First, the model is based on the assumption that a unidimensional latent trait is being measured and second, the model assumes local independence (i.e., for any given person, responses to one item in no way effect responses to another item). These last two assumptions are relatively easily met with careful test construction procedures.

Once the theoretical model had been decided upon and the assumptions had been evaluated to determine applicability of the model, the major problem became the actual implementation. More

specifically, estimation procedures needed to be determined for the ability and easiness parameters. Several techniques had already been developed for the estimation of the parameters on a paper and pencil test and these techniques were readily applicable to the calibration of items for use with the tailored testing program. The currently available techniques can be classified into three categories. First there is the original least squares "eyeball" approach that Rasch used in his original presentation of the model (Rasch, 1960). Second, Brooks (1964) used least square regression techniques to quantify the graphic techniques used by Rasch and finally, Panchapakesan (1969) developed a maximum likelihood technique that has been programmed for use on a computer (Wright and Panchapakesan, 1969).

Based on information given in Panchapakesan's dissertation (Panchapakesan, 1969), the maximum likelihood technique seems to yield superior results and hence was used for item calibration in this study. The actual computer program used for calibration was a greatly modified version of a program obtained from Jerry Durovic of the New York State Civil Service Department.

Since the ability parameters of the simple logistic model needed to be estimated in real time after each item had been administered to an individual, procedures developed for use with standard group tests were no longer appropriate. As an alternative an algebraic maximum likelihood solution was attempted, but solution of the necessary equations required finding the roots of high order polynomials and hence the algebraic procedure was dropped. Instead, a computer program was written that searched the likelihood function for its maximum. On trial runs, the program

was found to converge fairly rapidly on the maximum (approximately seven iterations required) and hence it was reasonable to use the procedure for real time estimation of ability. In practice, operating system time lags are much more noticeable than the time required to estimate ability parameters.

Along with ability parameter estimation, a procedure was also needed to determine a lower bound on the ability estimate to be used for classification purposes. Since the likelihood function was already being used to estimate ability, the area beneath the likelihood function was found using numerical integration and the lower 5% point of this distribution was set as the lower bound on ability. This procedure is equivalent to a Bayesian procedure assuming a rectangular prior with bounds from zero to one hundred.

A final theoretical problem required a solution before the tailored testing procedure could be implemented. That problem was how the items to be administered to an individual were to be chosen. Lord (1970) presented many possible schemes for item selection from a fixed stepsize, up and down method to variable stepsize Robbins-Munro process. Weiss & Betz (1973) have also presented an extensive summary of techniques for item selection.

The particular technique chosen for implementation involves first estimating a person's ability parameter and then picking an item for administration with easiness parameter greater than or equal to the reciprocal of the ability parameter. This procedure results in the selection of an item with a traditional difficulty index of 50 or easier. If no ability parameter estimate is available an item with easiness parameter 1.00 is selected and a fixed step procedure is used until both correct and incorrect

responses have been obtained. The procedure is discussed in greater detail in Reckase (1974).

III. Method

As described in the previous sections, the purpose of this paper is to present the results of a study into the practical application of tailored testing. In this section the actual data collection procedure will be described, including a description of the subject sample used. The results presented here are based on a pilot study for a more elaborate evaluation that is currently being planned. While the sample is small, the data generated give valuable information concerning the usefulness of the tailored testing model.

The sample used for this study was composed of seventeen Ss from a graduate-undergraduate measurement course at the University of Missouri who volunteered to participate in the experiment. The Ss ranged from college juniors to 2nd year graduate students and were approximately evenly divided between males and females.

During the experiment each S was evaluated on an individual basis in two ways. When an S arrived for the experimental session he was first administered a fifty item multiple-choice exam on statistics and measurement concepts. The test was administered in a small room, to minimize interruptions and distractions, without any time limitations. After completing the paper and pencil test, the S was taken to a second room containing an IBM 2741 typewriter terminal and signed on to an IBM 370/165 computer for the tailored testing procedure. The program accessed then typed out

instructions and quizzed the S for his student number and the subject code for the subject matter area to be tested. The E stayed with the S until all questions had been answered and it was clear that the S understood the operation of the terminal. The E then left the room and the administration became self-paced. The testing situation continued until a decision was reached or until the item pool was depleted. The instructions to the S and a sample item are shown in Figure 1.

The paper and pencil test used as a pretest was calibrated on 250 students in an undergraduate measurement course using the maximum likelihood program developed by Wright and Panchapakesan (1969) described in Section II of this paper. The subject matter area, statistics and measurement concepts, was chosen because the greatest number of items were available in the tailored testing item pool in that area. Forty statistics and measurement items had been stored in the tailored testing data set for use and were available for this study. The items in the item pool were calibrated on 250 to 966 students from an undergraduate measurement course over a period of two years. Details of the item storage format are given in Reckase (1974).

From the two testing situations described above the following data were gathered on each S including (1) the raw score on the paper and pencil test, (2) the corresponding ability estimate, (3) the letter grade classification for the subject on the test, (4) the final ability estimate based on the tailored testing procedure, (5) an estimate of the lower limit on the ability estimate, (6) the letter grade assigned, and (7) the number of items administered by the tailored testing procedure. These measures were then analyzed

to answer three questions. First, do the tailored testing procedure and the paper and pencil test yield comparable ability estimate: second, do the tailored testing procedure and the paper and pencil test classify the Ss in the same way as to letter grade: and third, how many items were needed by the tailored testing procedure to converge on an ability estimate?

In order to answer the experimental questions the following statistical procedures were used. To compare the ability estimates obtained by the two techniques, both the Pearson Product moment and the Spearman Rank correlation were computed since the scale properties of the ability scales are unclear. The similarity of classification was determined using Kendall's τ statistic (Siegel, 1956) and the number of items needed for convergence is shown by the distribution of the values and summary statistics including the mean and standard deviation. These results with the raw data are presented in the next section.

IV. Results

The data for the analysis comparing the ability estimates obtained using each of the methods are shown in Table 1 along with the other measures obtained in the study. The Pearson Product Moment correlation between the two sets of ability estimates is 0.61 and Spearman rank order correlation is 0.73. If the ability estimates are on a ratio or interval scale, the former value is more appropriate, if the scale assumptions are not met the rank order is more appropriate. In order to interpret these correlations, the reliabilities of each of the procedures is

required. The KR-20 reliability of the paper and pencil test is available and is 0.72, but no data is available on the reliability of the tailored testing procedure. Both of the correlation coefficients are significant beyond the 0.05 level. Summary statistics on the ability estimates including the mean, median and standard deviation are given in Table 2. There are no significant differences between any of these statistics.

The similarity of the grade classifications of the two methods is summarized in the two-way table given in Table 3. A τ statistic showing similarity has been computed on this data yielding a value of 0.57. This statistic is significantly different from zero beyond the 0.001 level.

The results concerning the number of items required to classify a S into a grade category are shown in Table 4. Given are a frequency distribution for the number of items needed and the following descriptive statistics: the mean, median, mode, standard deviation, and range. From these data it can be seen that the distribution is highly positively skewed with a median value of twelve. This should be compared to the fifty items used on the paper and pencil test.

V. Discussion

Interpreting the results of this study is somewhat problematic because it is difficult to decide what results are desireable. Should the tailored testing procedure ideally yield ability estimates and grade classification identical to those obtained by the less than perfect paper and pencil test or should the ability

estimate be different, reflecting a perhaps "better" tailored testing procedure? It seems that at the very least there should be some similarity between the methods since they are trying to do the same thing; but if the tailored testing procedure is more accurate, the similarity should not be too great in light of 0.72 reliability of the paper and pencil test.

The correlational data obtained in this study show that the two methods yield similar results, but that the ability estimates are far from equal and the grade classifications are the same for only two-thirds of the Ss. This leaves open the possibility that the tailored testing procedure is an improvement over the traditional test, but needless to say, several other possibilities are available to explain the moderate correlations.

First, the item administration program terminates the testing session once a grade of A has been obtained. This occurred after as few as six items had been administered, which is hardly an adequate number for good estimation. If the administration of items had been allowed to continue after the assignment of an A grade, more accurate estimates would probably have been obtained and the agreement of the estimates with those of the paper and pencil test might have been better.

A second source of error in the estimation of ability is in the number and quality of items in the item pool. The item pool used for this study contained only forty items, some of which were of poor quality. Recent simulation studies seem to indicate that about 250 items are required for good estimation. If the item pool is small, the simulations show that the procedures used in this

study will tend to overestimate ability. In light of the simulated results, the forty item pool seems to have done amazingly well.

Another difficulty related specifically to the grading procedure is the technique for estimating the lower limit of ability. The lower 5% point on the likelihood distribution is at best a rough indication of lower limit. Bayesian procedures based on various different families of prior distributions are currently being studied as alternatives.

Although the procedure is beset with the problems just described, the end result has for the most part been positive. The procedure has been shown to work and most of the problems discussed can be overcome by reprogramming and by increasing the size of the item pool.

A more positive outcome of this study is the determination of the number of items required to classify Ss into grade categories. As is shown in Table 4, the distribution is highly positively skewed with a median value of twelve items. This is a substantial reduction from the fifty items in the standard test, although administration of the test is slower since each item is typed out during the testing session. The time needed for the administration of thirty tailored items is about equal to the time needed for the fifty item paper and pencil test. The use of faster cathod ray terminals will greatly improve test administration time.

In summary, the tailored testing procedure described in this paper has been shown to yield similar, but not equivalent, results to those of a conventional test in both the estimation of ability parameters and the asignment of letter grades. These results are obtained using substantially fewer items than the conventional test

while administering different items to each individual. Problems encountered in the operation of the procedure were also discussed including the size of the item pool, limits on the ability estimates, and problems with the stopping rule. Overall, the procedure has been shown to be a viable tailored testing method, worthy of further research and refinement.

Figure 1

```
TERMINAL TESTING PROCEDURE
YOU WILL BE PRESENTED WITH A SERIES OF TEST ITEMS.  RESPOND TO
EACH ITEM BY TYPING THE APPROPRIATE LETTER AND PRESSING THE
RETURN KEY.  ITEMS WILL BE PRESENTED UNTIL A CLEAR DECISION
IS REACHED CONCERNING WHETHER YOU ARE ABOVE OR BELOW A C GRADE.
IF YOU WISH TO CONTINUE ON FOR A HIGHER GRADE, INSTRUCTIONS
WILL BE GIVEN AT THAT POINT.  IF AT ANY TIME YOU WISH TO STOP
BEFORE A DECISION HAS BEEN MADE, TYPE THE WORD  STOP  AFTER
YOUR LETTER RESPONSE AND PRESS THE RETURN KEY.


PLEASE TYPE YOUR STUDENT NUMBER AND PRESS THE RETURN KEY
IF YOUR STUDENT NUMBER CONTAINS ONLY 5 DIGITS
START IT WITH A LEADING ZERO TO MAKE SIX DIGITS


100000
INPUT:  ID = 100000
TYPE THE CODE CORRESPONDING TO THE AREA YOU ARE TO BE TESTED ON
          SM FOR STATISTICS AND MEASUREMENT
          ET FOR CLASSROOM EVALUATION TECHNIQUES
          ST FOR STANDARDIZED TESTS
AFTER TYPING THE PROPER CODE, PRESS THE RETURN KEY
sm
INPUT: TEST CODE = SM


      1
          A PSYCHOLOGIST WHO WANTS A MEASURE OF THE EXTENT TO WHICH
  SCORES IN A GROUP VARY MIGHT CONCEIVABLY CHOOSE ANY ONE OF THE
  FOLLOWING EXCEPT
                (A) THE RANGE.
                (B) THE VARIANCE.
                (C) THE STANDARD DEVIATION.
                (D) THE MEDIAN.
TYPE RESPONSE LETTER AND PRESS RETURN


d
CORRECT
```

Table 1

Paper and Pencil and Tailored Test Raw Data

	Paper and Pencil Test			Tailored Test			
Subject	Raw Score	Ability Estimate	Letter Grade	Ability Estimate	Lower Limit	Letter Grade	Number of Items
1.	38	4.657	B	18.452	4.25	A	6
2.	38	4.657	B	18.452	4.25	A	6
3.	33	2.417	C	2.900	1.50	C	34
4.	28	1.360	-	0.213	0.10	-	13
5.	36	3.530	C	0.627	0.41	-	20
6.	38	4.657	B	1.842	0.97	C	18
7.	44	13.718	A	3.803	1.94	B	39
8.	46	23.786	A	32.470	6.91	A	12
9.	43	10.983	A	8.745	3.49	A	14
10.	47	34.289	A	35.096	7.55	A	10
11.	41	7.494	A	17.766	4.11	A	8
12.	48	55.866	A	18.452	4.25	A	6
13.	46	23.786	A	18.452	4.25	A	6
14.	31	1.909	C	1.726	0.55	C	21
15.	42	8.994	A	11.516	3.52	A	9
16.	40	6.330	A	11.516	3.52	A	9
17.	40	6.330	A	1.929	1.03	C	28

Table 2

Summary Statistics
on Ability Estimates

	Paper and Pencil Test	Tailored Test
Mean	12.62	11.99
Median	6.330	11.516
Standard Deviation	13.00	10.55
Reliability KR-20	0.72	---
Correlation		
Pearson Product Moment	0.61*	
Spearman ρ	0.73**	

* $p < 0.05$
** $p < 0.01$

Table 3

Similarity of Grade Classification

		Tailored Test Classification			
		A	B	C	D
Paper and Pencil Classification	A	0	1	1	0
	B	2	0	1	0
	C	0	0	2	1
	D	0	0	0	1

Kendall's τ = 0.57***

*** $p < 0.001$

Table 4

Summary of the Number of Items Required for Tailored Testing Procedure

Frequency Distribution

Number of Items Required	Frequency
1 - 5	
6 - 10	8
11 - 15	3
16 - 20	1
21 - 25	1
26 - 30	2
31 - 35	1
36 - 40	1

N = 17

Mean = 15.59

Median = 12.00

Mode = 6.00

Standard Deviation = 10.14

Range = 6-39

References

Birnbaum, A. Some latent trait models and their use in inferring an examinees ability. In F.M. Lord and M.R. Novick, Statistical Theories of Mental Test Scores, Reading, Mass: Addison-Wesley, 1968, Chapters 17-20.

Brooks, R.D. An Empirical Investigation of the Rasch Ratio-scale Model for Item-Difficulty Indexes. Doctoral Dissertation, University of Iowa, 1964. No. 65-434.

Hambleton, R.K. An Empirical Investigation of the Rasch Test Theory Model. Doctoral Dissertation, University of Toronto, 1969.

Linn, R.L., Rock, D.A. & Cleary, T.A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Lord, F.M. Some test theory for tailored testing. In W.H. Holtzman (Ed.), Computer-assisted Instruction, Testing and Guidance, New York: Harper & Row, 1970.

Panchapakesan, N. The Simple Logistic Model and Mental Measurement. Doctoral Dissertation, University of Chicago, 1969.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Denmarks Paedagagiske Institut, 1960.

Reckase, M.D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods and Instrumentation, in press.

Ross, J. An empirical study of a logistic mental test model. Psychometrika, 1966, 31, 325-340.

Siegel, S. Nonparametric Statistics, McGraw-Hill, New York, 1956.

Weiss, D.J. and Betz, N.E. Ability measurement: conventional or adaptive? Research Report 73-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, February, 1973.

Wright, B.D. and Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.